# Context Awareness and Priority Control for ITS based on Automatic Speech Recognition

Sakriani Sakti*‡, Oyunchimeg Shagdar‡, Fawzi Nashashibi‡, and Satoshi Nakamura*
*Graduate School of Information Science, Nara Institute of Science and Technology, JAPAN
Email: {ssakti, s-nakamura}@is.naist.jp
†RITS Project-Team, INRIA Rocquencourt, FRANCE
Email: {oyunchimeg.shagdar, fawzi.nashashibi}@inria.fr

***Abstract*—Bringing rapid assistance to motorists involved in a traffic accident is an important service to be provided by Intelligent Transportation System (ITS). Existing proposals to automatic accident detection are based on the vehicle's perception point of view. This paper introduces situational awareness based on the "understanding" of conversational speech of drivers/passengers using an automatic speech recognition (ASR) system. Context-aware priority control and congestion control schemes are presented to ensure coexistence of ASR-triggered applications and cooperative awareness messages (CAM) in the IEEE 802.11p system. Finally, application risk analysis and performance evaluations of ASR and V2X communications are carried out.**

## I. Introduction

The European Parliament launched a so-called emergency call (eCall) initiative with the purpose to bring rapid assistance to motorists involved in a collision anywhere in the European Union. In the same context, the European Telecommunications Standardisation Institute (ETSI) defined the SOS service, in which an SOS alarm is transmitted to a service centre in case of life threatening emergency [1]. The service delay can be significantly short if the accident is detected automatically based on e.g., a horizontal tilt sensor embedded in the vehicle. However, if the car fails to automatically detect the situation, the driver/passengers have to call an emergency service and follow the existing standard protocol answering a series of questions that may require a long delay until an ambulance is dispatched.

What would be a breakthrough, in the authors' opinion, is that the emergency level shall be automatically detected not only from the vehicle's point of view but also from the behaviours of the drivers/passengers, who are in the vehicles, which are involved in the accident, or in the nearby vehicles. This paper reports our study on automatic detection and notification of emergency and other road situations based on "understanding" of conversational speech of drivers/passengers using an ASR system. In fact, the use of ASR that automatically recognizes drivers' command words has been studied for many years in order to allow drivers to give command and control through speech (i.e., play music, ask navigation, etc) while keeping their eyes and hands focused on driving. The Verbmobil was first defined in 1992 with an eight years initiative project to apply multilingual speech and language technology inside cars [2]. However, the majority of the existing efforts on ASR concentrate on developing a robust in-vehicle ASR system [3], [4], [5]. In this paper, we aim to utilise ASR not only for drivers' comfort but also for emergency assistance. More specifically, ASRs are designed to be embedded in vehicles that constantly recognize the drivers'/passengers' conversational speech, and classifies the spoken data into different levels of importance (i.e., daily conversation or SOS requests). Then, based on the context, a vehicle to Internet communication is triggered to call the necessary service.

Since the IEEE 802.11p [6] (and the European standard: ETSI ITS-G5 [7]) is developed to support various types of vehicular applications [1], it seems logic to study its applicability to the above-mentioned ASR-triggered communication. The IEEE 802.11p has been the focus of a great number of R&D activities, and its applicability to road safety and efficiency applications have been tested in some projects [8]. The key weakness of the IEEE 802.11p is the channel congestion problem, where channel is saturated when the number of the 802.11p-equipped vehicles is large. This problem is obviously due to the limited resource at the 5.9 GHz band, but also because all the vehicles are expected to periodically broadcast CAMs, which are needed for collision avoidance but tend to load the wireless channel. Several distributed congestion control (DCC) algorithms [1], [9] are proposed as a result of the recent efforts. In this paper, we are interested in studying the impacts of the IEEE 802.11 priority control and DCC for coexistence of ASR-messages and CAMs.

The main contributions of this paper are as follows.

- Propose context aware priority control: The priority level of a data traffic is generally fixed based on the application type (i.e., video streaming uses VI, email uses BE). In contrast, this work enables priority setting based on the importance or urgency of the content of speech message, in which the "understanding" of speech message is done automatically using ASR.
- Study the impacts of priority control in channel congestion situation: Priority control is designed to enable higher communication performance to the data belonging to higher priority class. However, this study finds that the priority control mechanism provides a very little impact when wireless channel is congested, and therefore we suggest to apply priority control jointly with a DCC in order to achieve an satisfying communication performance.

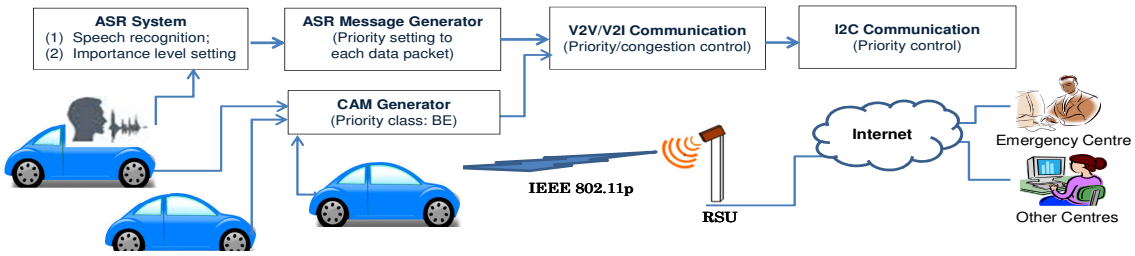



Fig. 1: *System overview.*

## II. ASR Enabled ITS Services using V2X Communications Systems

### A. Overall System Design

The target system is illustrated in Fig. 1. As specified by the ETSI ITS standards, all vehicles periodically generate CAM packets, which contain the position and moving direction of the vehicles. Vehicles are also equipped with ASR systems that constantly recognises speech drivers/passengers conversational speech and then classifies the spoken data into different levels of importance. Here, we defined four level of importance: (1) basic conversation such as hotel or restaurant reservation; (2) road-related conversation such as traffic-jam, navigation, etc; (3) sickness-related conversation that require to contact doctor or medicament; (4) accident-related conversation that may require an emergency call. After that, it creates data packets (we call ASR message) containing conversation messages and the tag of the importance level. This tag classification of conversation message correspond to the ITS applications of (1) point of interest (POI), (2) traffic efficiency (3) remote medical support and diagnosis, and (4) SOS, respectively. An ASR message is then sent to a RSU, which will further send the message to the target destination server over the Internet. In the following subsections, we detail the ASR and V2X communication systems.

### B. ASR System

Given feature vectors of the speech signals $S_S = [x_1, x_2, ..., x_T]$, the state-of-the-art of statistical speech recognition task is to find a word sequence $W_S = [w_1, w_2, ..., w_N]$ that maximizes the conditional probability $P(W_S|S_S)$:

$$\hat{W_S} = \arg\max_{W_S} P(W_S|S_S) = \arg\max_{W_S} P(S_S|W_S)P(W_S). \quad (1)$$

Here, $P(W_S)$ is called a *language model* ($LM$) and represents a priori probability for the word sequence independent of the speech input signal, while $P(S_S|W_S)$ is called an *acoustic model* ($AM$) and represents the likelihood that the speech signals of the source language were generated by the model.

The speech features are extracted every 10 ms with 25 ms width using a widely known front-end mel-frequency cepstral coefficients (MFCC). To incorporate the temporal structures and dependencies, several adjacent frames of MFCCs are stacked into one single super vector, which then reduced to an optimum dimensions by applying a linear discriminant analysis (LDA). After that, the resulting features are further de-correlated using maximum likelihood linear transformation (MLLT) [10], which is also known as global semi-tied covariance (STC) [11] transform. Moreover, speaker adaptive training (SAT) [12] is performed using a single feature-space maximum likelihood linear regression (fMLLR) [11] transform estimated per speaker.

Acoustic models are trained on the features describe above. Here, we applied two kinds of acoustic models: (1) Hidden Markov Model/Gaussian Mixture Model (HMM/GMM) which is a standard context-dependent cross-word triphone with a three-state HMM topology. The HMM units are derived from 39 phonemes of English, and they were trained with GMM output probability; and (2) Hidden Markov Model/Deep Neural Network (HMM/DNN) in which DNN replace the GMM output probability. DNN used here is based on generalized maxout networks that uses non-linearity dimension-reducing of p-norm (adopted from [13]). For language model, we applied a standard n-grams language modelling approach, where $N-1$ words are used as context to predict the next word. Here, we built trigram language models with Witten Bell smoothing [14]. Our decoding algorithms use weighted finite state transducers (WFSTs) [15].

Based on ASR outputs, we classifies the content into four importance levels as described in Section II-A. Here, classification is done simply based on a predefined critical words. In consequence, if those critical words are failed to be recognized, then ASR may put the data to a wrong importance level.

### C. V2X Communications System

Solutions including IP mobility management and fast handover [16] are necessary for delivering data from a vehicle to a server, which is in the Internet. The main interest of this work is, however, coexistence of ASR messages, which belong to different classes, and CAMs. ASR messages tagged with different importance level by the ASR system require transmissions to different service centres: POI, traffic efficiency, medical support, and SOS. We assume that the server addresses are known at the vehicles thank to service announcements made by RSUs [1]. Due to the different quality requirements of different classes, it seems necessary to consider prioritisation to ASR messages and CAMs. Note that, although, we can expect to transmit the packets of e.g., SOS in the control channel (CCH) and the packets of e.g., POI in the service channel (SCH), due to the possibility of vehicles being equipped with single interfaces, we consider transmissions of ASR messages and CAMs in the same frequency channel.

Since wireless channel is much more lossy than that of the wired systems, we focus on the V2X wireless communication with priority and congestion control schemes:

- **Priority control**
  The IEEE 802.11p provides differentiated channel access to four access classes (ACs): VO, VI, BE, and BK, which operate with different settings to Contention Window size and Arbitrary Inter-Frame Space [6]. However, in the existing communication systems, packet classification to different ACs is made in a static manner often based on the application type (e.g., video streaming uses VI class and email uses BE). In contrast, we present two priority control strategies that exploit the context awareness of the ASR system:
  1) *Adaptive priority control:* Each ASR message with four different important levels is mapped directly to the four ACs in their order of the importance (i.e., accident-related is mapped to V0, basic conversation is mapped to BK). The decision is done in each message basis.
  2) *High-class priority control:* If one ASR message is considered to possibly include an SOS alarm, then all ASR messages will belong to the highest priority level: VO. The decision is at one time for all messages.

  In any of the above-cases, CAMs are transmitted in BE.
- **Congestion control**
  Since ASR messages share the channel resource with CAMs, even if they sent on a high-priority AC, we believe that it is important to consider channel congestion. ETSI introduced Reactive DCC algorithm, in which each node monitors the channel load (CL: channel busyness ratio) and controls the CAM rate following a parameter table, which maps CL and CAM generation interval [7]. In [9], [17], it is shown that Reactive DCC creates channel load oscillation, which largely degrades the communication performance. In [17], we introduced a simple extension of Reactive DCC, Asynchronous Reactive DCC, which stabilises the channel load state based on randomised interval setting. To this end, in this paper, we apply Asynchronous Reactive DCC to CAMs and evaluate its impact on the V2X communication performance.

### D. ITS Services

To measure the reliability of ASR-based ITS services, we define the overall expected (average) loss function that measures losses incurred in transmitting a message based on both ASR and V2X communication systems:

$$\mathbb{E}[L] = \sum_{k} \sum_{j} \int_{R_j} L_{kj}[P_{ASR}(x, C_k) P_{V2X}(x)]dx, \qquad (2)$$

where $L_{kj}$ is a loss function of an action that recognize and transmit speech message $x$ with importance level $j$, while in the truth condition $C_k$ speech message $x$ belongs to importance level $k$. $\int_{R_j} P_{ASR}(x, C_k)$ is the error probability, where ASR classifies $C_k$ into the decision region $R_j$. $\int_{R_j} P_{V2X}(x)$ is the transmission error of x, which has the importance level of $j$ (we assume that ASR and V2X errors are independent).

Obviously, $L_{kj} = 0$ if $k = j$. Now we consider two types of critical errors: (1) "False positive" in which a non emergency message (e.g., basic daily conversation) is incorrectly recognized by the ASR as an emergency message and transmitted by the communication system to the SOS server. In this case, an SOS service may be activated but not used; (2) "False negative" in which an SOS request is incorrectly recognized by the ASR as a non-emergency conversation, or it is correctly recognized but the communication system failed to transmit to the SOS server. In this case, no emergency service is provided to the drivers/passengers, who need a treatment, resulting in a possible death. The consequences of the above-mentioned two errors are dramatically different; the loss function of the false negative should be set to a significantly larger value compared to that of the false positive.

## III. Performance Evaluation

### A. ASR Performance

We applied the state-of-the-art ASR technologies described in Section II-B based on Kaldi speech recognition toolkit [18]. Our original ASR engine [19] was basically trained for general purpose using TED-LIUM corpus [20]. It is a speech data corpus made from open-domain spontaneous speech of TED Talks[1], including audio talks and their transcriptions available on the TED website. In total, there are 774 audio talks which consists of about 118 speech hours. For language model, we use cantab Language model [21] consisting 155M tokens entropy from 150K word vocabulary that extracted from 1 billion word of google n-gram[2]. We call these models as "base-AM" and "base-LM", respectively.

As topic-dependent modeling has proven to be an effective way of improving the quality of models, we also utilize the ATR basic travel expression corpus (BTEC) that has served as the primary source for developing broad coverage speech translation systems [22]. The sentences were collected by bilingual travel experts from Japanese/English sentence pairs in travel domain phrasebooks that covers travel conversation, including medical sickness or accident related conversation. In total, there are about 400 speakers where each speaker utter about 400 sentences. It includes also several accents of English: United States, Australian, and British. We call these models as "adapt-AM" and "adapt-LM", respectively.

The test data is selected from BTEC test set which consist of 2400 utterances (there are about 8 speakers where each speaker utter about 300 sentences) covering four level of importance described in Section II. The sentence examples of each level of importance are shown in Table I. To investigate the impact of the ASR quality on ITS applications, we simulate various feature extractions, AM and LM set-up as described in Table II. The ASR word accuracy and the tag classifier accuracy are shown in Fig. 2. As expected, the better speech

[1] http://www.ted.com/talks

[2] https://code.google.com/p/1-billion-word-language-modeling-benchmark/

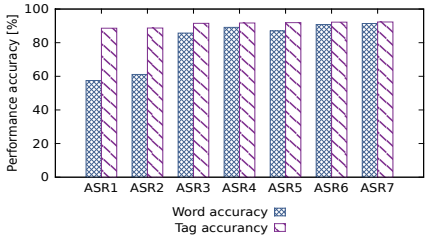
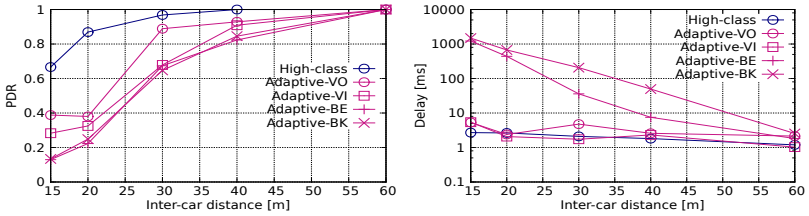
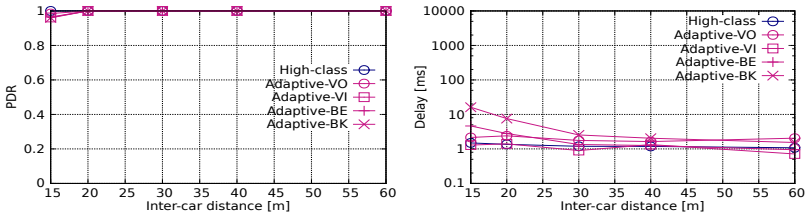



Fig. 2: *ASR performance based on word accuracy and tagging accuracy.*




(a) PDR (b) Delay

Fig. 3: *Performances of prioritised V2X communication without DCC.*




(a) PDR (b) Delay

Fig. 4: *Performances of prioritised V2X communication with DCC.*

TABLE I: Example of ASR messages on 4 priority level.

| **L**evel | **A**SR messages |
|---|---|
| 1 | How much does it cost to get to the restaurant by taxi? |
| 2 | How many hours before we arrive in Tokyo? |
| 3 | I feel sick. |
| 4 | There's been a traffic accident. Someone has been injured. |

TABLE II: Simulation of various quality of ASR systems.

| **A**SR systems | **F**eat | **A**M | **L**M |
|---|---|---|---|
| ASR1 | MFCC-deltas | Base-AM | Base-LM |
| ASR2 | LDA-MLLT | Base-AM | Base-LM |
| ASR3 | LDA-MLLT | Base-AM | Adapt-LM |
| ASR4 | LDA-MLLT | Base-AM (DNN) | Adapt-LM |
| ASR5 | LDA-MLLT-SAT | Base-AM | Adapt-LM |
| ASR6 | LDA-MLLT-SAT | Base-AM (DNN) | Adapt-LM |
| ASR7 | LDA-MLLT-SAT | Adapt-AM (DNN) | Adapt-LM |

feature representation is used, the better the ASR accuracy, and the topic-adapted model gives significant improvement compared to general purpose model (base-AM and base-LM). Nevertheless, the AM1 with 57.5% word accuracy is surprisingly able to reach 88.6% tag accuracy. However, in the current evaluations, the errors are considered as equivalent. In Section III-C, we will discuss risk analysis, where the errors have different impacts on ITS services.

### *B. V2X Communication Performances*

The packet delivery ratio (PDR) and end-to-end delay performances of the ASR traffic are evaluated using the NS3 network simulator (version 3.21), which includes the TCP/IP protocol stuck and the IEEE 802.11p. CAM and ASR message generators and the DCC algorithm are implemented to the NS3. Simulations are carried out for a scenario and parameters suggested by ETSI [7]. Specifically, vehicles are uniformly distributed on a highway with length of 1000 meters. RSUs are installed in the middle line (between two directions). If no DCC is applied, the CAM generation interval is 0.1 seconds. The rest of the simulation parameters are listed in Table III. The ASR system (see Fig.1) converts spoken data to text

TABLE III: Simulation parameters for V2X communication.

| | |
|---|---|
| Simulation time | 300 seconds |
| Highway structure | 3-lanes/2-directions |
| Inter-car distance | [15, 60] m |
| RSU inter-location | 200 m |
| Fading model | LogDisance, exponent 2 |
| Wireless access technology | IEEE 802.11p (6Mbps) |
| Transmission power | 23 dBm |
| Modulation scheme | QPSK 1/2 |
| ASR/CAM data size | 200/400 Bytes |
| ASR packet interval | 5 seconds |

data and assigns importance levels to each text (sentence). The ASR message generator implemented in NS3 reads the text files, generates packets, and sends with priority labels following *High-class* or *Adaptive* priority control strategies (See Section II-C). 1 out of 10 vehicles are randomly selected to create an ASR data that consists of 50 packets (sentences). The ASR messages are routed to the service centres via the RSU, which corresponds to the strongest receive signal strength.

Figures 3 and 4 are the simulation results for different inter-car distances obtained without and with DCC, respectively. Note that due to the space constraint, we do not show CAM results. The detailed evaluations on CAM performances under DCC can be found in [17].

From Fig. 3, we first notice rather obvious results for *Adaptive* priority control: the packets sent at higher priority achieve better performance. On the other hand, what is less obvious is that *High-class* priority control provides always better performance than any of the ACs of *Adaptive* strategy, even that of the VO class (Adaptive-VO). The reason can be explained as follows. In *High-class* strategy, all ASR messages are sent in the VO class, thus only V0 and BE (CAMs) classes compete for channel access and hence the VO (ASR) packets get a very good change to succeed thank to the large differentiation between VO and BE classes [6]. On the other

hand, in *Adaptive* strategy, since ASR packets are classified to 4 ACs, all the ACs participate to competition, reducing the success probability of the individual ACs due to the small differentiation between neighbouring ACs (BK and BE, BE and VI, and VI and VO).

Figure 4 compare the performances when DCC is applied. The figure clearly shows the importance of congestion control: the performances of all the ACs is sufficiently high regardless of the priority control strategies.

### C. Risk Analysis of ITS Services

Figure 5 depicts the overall expected loss given the worse and the best ASR systems (ASR1 and ASR7) for the scenarios where the inter-car distance is 15, 30, and 60 m, and with or without DCC. As can be expected, the denser the traffic is the higher the expected loss regardless the performance of ASR system. Considering ASR performance, the consequences of two possible ASR errors (described in Section II-D) are dramatically different. Hence the loss penalty set to the false negative is 1000 times greater than that set to the false positive (see Section II-D). Figure 5 reveals that the significant difference in word accuracy give a significant impact in the expected loss. ASR1 (base-AM and base-LM) with 57.5% word accuracy provide very high risk, conceivably because it failed to recognise many critical words. The optimum results are provided by ASR7 (adapted-AM and adapted-LM) with 92.3% word accuracy which can provide nearly zero risk when priority control and DCC are performed.

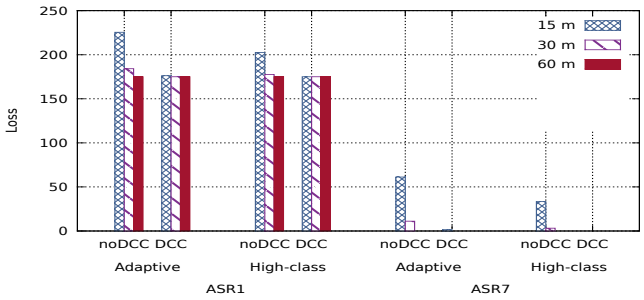



Fig. 5: *Expected loss of ITS applications*

## IV. Conclusion

This paper presents our work on context-awareness and priority control in V2X communications exploiting automatic speech recognition (ASR) system. Drivers/passengers conversation are converted to text data and classified to different access categories (ACs) based on the importance of the content. Coexistence issue of ASR traffic and CAMs is studied; in order to ensure the quality of the ASR traffic, context-aware priority control and distributed congestion control (DCC) schemes are jointly applied. The simulation results reveal the feasibility of context-aware priority control as well as the importance of DCC. The application risk analysis shows that ASR7 (adapted-AM and adapted-LM) scheme, which has 92.3% word accuracy, can provide nearly zero risk, when a priority control and DCC are applied for the V2X communication.

However the present studies were confined to relatively controlled data sets. In real condition, drivers/passengers may only chat about SOS among themselves, without an intention to request an ITS service. Therefore, in our future work, it would worthwhile to investigate robust classification techniques to define importance level of ASR messages, that could consider and differentiate drivers/passengers intention.


## Acknowledgment

Part of this work was executed under JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation.


## References


[1] *ETSI TR 102 638; Intelligent Transport Systems (ITS); Vehicular communications; Basic set of Applications*, Std., Feb. 2009, v1.0.4.

[2] W. Wahlster, *Verbmobil: Foundations of speech-to-speech translation*. Springer, 2000.

[3] J. Hernando and C. Nadeu, "Speech recognition in noisy car environment based on OSALPC representation and robust similarity measuring techniques," in *Proc. ICASSP*, 1994, pp. 69–72.

[4] W. Li and H. Bourlard, "Sub-band based log-energy and its dynamic range stretching for robust in-car speech recognition," in *Proc. INTERSPEECH*, 2012, pp. 314–317.

[5] X. Feng, B. Richardson, S. Amman, and J. Glass, "On using heterogeneous data for verhicle-based speech recognition a DNN-based approach," in *Proc. ICASSP*, 2015, pp. 4385–4389.

[6] *IEEE Standard for Information technology — Telecommunications and information exchange between systems — Local and metropolitan area networks — Specific requirement, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE Computer Society Std., July 2010, iEEE Std 802.11p-2010.

[7] *ETSI TR 101 612; Intelligent Transport Systems (ITS); Cross Layer DCC Management Entity for operation in the ITS G5A and ITS G5B medium; Report on Cross layer DCC algorithms and performance evaluation*, Std., Sep. 2014, v1.1.1.

[8] Drive-C2X, "thttp://www.drive-c2x.eu/publications."

[9] J. B. Kenney, G. Bansal, and C. E. Rohrs, "LIMERIC: a linear message rate control algorithm for vehicular DSRC systems," in *Proceedings of the Eighth ACM international workshop on Vehicular inter-networking*. ACM, 2011, pp. 21–30.

[10] R. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," in *Proc. of ICASSP*, 1998, pp. 661–664.

[11] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[12] R. S. T. Anastasakos, J. Mcdonough and J. Makhoul, "A compact model for speaker adaptive training," in *Proc. ICSLP*, 1996, pp. 1137–1140.

[13] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *Proc. of ICASSP*, Florence, Italy, 2014, pp. 215–219.

[14] I.-H. Witten and T.-C. Bell, "The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression." *IEEE Transactions on Information Theory*, vol. 37, pp. 1085–1094, 1991.

[15] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech and Language*, vol. 20, no. 1, pp. 69–88, 2002.

[16] H. Yokota, K. Chowdhury, R. Koodli, B. Patil, and F. Xia, "Fast handovers for proxy mobile IPv6," Tech. Rep., 2010.

[17] O. Shagdar, "Evaluation of Synchronous and Asynchronous Reactive Distributed Congestion Control Algorithms for the ITS G5 Vehicular Systems," *INRIA Technical Report, TR 432*, April 2015, tR: 432.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Moticek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, Hawaii, USA, 2011.

[19] S. Sakti, K. Kubo, G. Neubig, T. Toda, and S. Nakamura, "The NAIST english speech recognition system for IWSLT 2013," in *Proc. of IWSLT*, Heidelberg, Germany, 2013, pp. 269–272.

[20] A. Rousseau, P. Delglise, and Y. Estve, "TED-LIUM: an automatic speech recognition dedicated corpus," in *Proc. of LREC*, Istanbul, Turkey, 2012, pp. 125–129.

[21] W. Williams, N. Prasad, D. Mrva, T. Ash, and T. Robinson, "Scaling recurrent neural network language models," in *Proc. of ICASSP*, Brisbane, Australia, 2015.

[22] G. Kikui, S. Yamamoto, T. Takezawa, and E. Sumita, "Comparative study on corpora for speech translation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1674–1682, 2006.